SYNTHESIS OF SHAKING VIDEO USING MOTION CAPTURE DATA AND DYNAMIC 3D SCENE MODELING

Shao-Ping Lu[1,2]*, Jie You*[1]*, Beerend Ceulemans*[1]*, Miao Wang*[3]*, Adrian Munteanu*[1]

[1]Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB)
[2]CCCE&CS, Nankai University [3]TNList, Tsinghua University

ABSTRACT

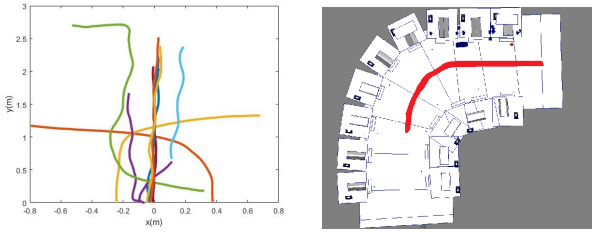
Important video processing methods such as video stabilization and deblurring often do not have ground-truth data available. This poses a great challenge in the development and parameter tunning of such methods. Synthetic shaken video is very useful to generate well-defined ground-truth datasets. Existing shaking video synthesis methods simulate shaky camera motion by performing 2D view warping using only a single 2D video, which does not always correspond to realistic 3D motions. In this paper, we introduce a novel shaking video synthesis approach. The proposed framework constructs the camera motion trajectory by making use of human motion information that is captured in the real-world. Moreover, we render the shaken video from man-made dynamic 3D scenes with detailed camera pose information. Our novel approach provides both accurate 2D visual content and camera motion trajectory in the 3D scene, which allows for evaluating the visual distortion as well as the offsets of the recovered camera trajectory. The proposed synthesis method of shaking video will benefit and ease future research on 3D-aware video stabilization.

Index Terms— Shaking video generation, dynamic 3D, motion capture, camera trajectory, video rendering

1. INTRODUCTION

With the increasing ubiquity of digital consumer cameras in the industry and for the general public, new requirements are emerging in the use of the captured visual content: interactive editing and modeling became a must for various high-quality video applications. In this context, many post-processing algorithms, such as video stabilization [1] and deblurring, have been widely investigated as means to satisfying the requirements of high-quality video applications.

In addition to a number of existing video stabilization algorithms, some recent efforts have been devoted to the design of shaking videos [2, 3, 4, 5]. The motivation behind such methods is that a well-defined benchmark (ground-truth) is of critical importance in this field. However, existing video stabilization methods take a smooth 2D video as input and generate an output video mainly by frame-level view warping. These methods lack accurate 3D information and thus cannot yield believable 6 degrees-of-freedom (DOF) camera motion. Although benchmarks for static 3D reconstruction have already been introduced recently [6], there is still a lack of stabilization datasets for *dynamic* 3D scenes.



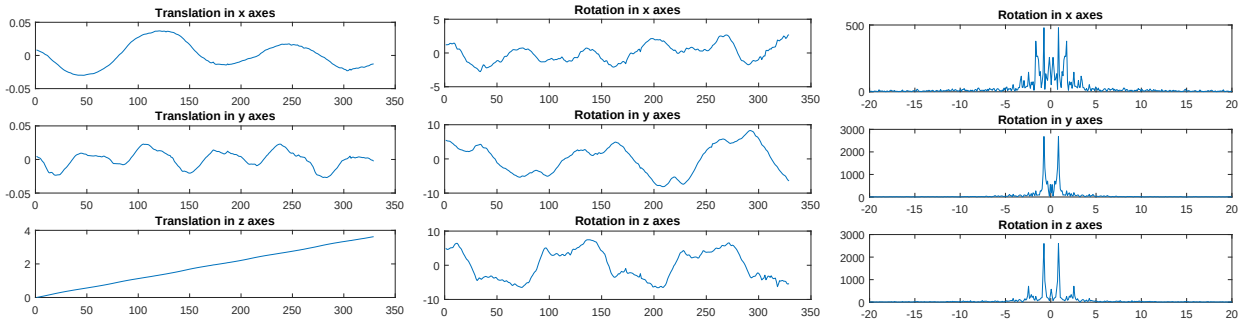
(a) Motion activity (b) Desired camera path

Fig. 1. Using human motion capture dataset (the left figure) and an expected motion path coarsely specified by the user (the red line in the right figure), our system aims to synthesize a desired camera trajectory and render the corresponding shaking/stable videos from the dynamic 3D scene.

This paper introduces a novel approach for the synthesis of shaking videos. The key insight is that we synthesize a shaking video from a known 3D scene rather than an existing 2D video. In order to achieve this, we take into account both human motion capture data and accurate 3D scene modeling. Nowadays, motion capture systems are capable to precisely measure human motion activities with multi-modal sensors, and they have already been widely used in commercial solutions [7, 8, 9] as well as various research areas [10, 11]. By introducing a motion capture dataset [12] into our system, we are able to provide a 3D shaking trajectory with detailed 6 DOF that are recorded from the real-world.

We argue that the availability of accurate 3D motion trajectories does not necessarily help to generate the desired shaking

The first two authors contributed equally. This work was supported by the Fund for Scientific Research-Flanders (FWO-Flanders) projects G084117 and G025615. Miao Wang is supported by National Natural Science Foundation of China (project number 61561146393) and China Postdoctoral Science Foundation (project number 2016M601032).



(a) Translation (b) Rotation (c) Spectrum of the rotation

Fig. 2. Analysis of the motion capture data obtained from [12]. The motion information is extracted from multi-modal sensors on the human body. In our solution we mainly focus on the 6 DOF, i.e. 3 translation components (the left column) and 3 rotation components (the middle column) on the 3D world coordinates, respectively. Note that the motion information can be considered periodic and its signal could thus be temporally extended using frequency-based reconstruction (see the spectrum magnitude in the right column).

videos when directly warping a single 2D video. As can be seen in Fig. 1, our system can generate a camera motion trajectory according to an expected motion path coarsely specified by the user, and meanwhile the detailed 6 DOF are carefully preserved from the motion capture dataset. Therefore, when rendering the output shaking video from a dynamic 3D scene, we can obtain much more accurate visual information than warping a 2D video.

In summary, to the best of our knowledge, we are the first to synthesize shaking videos by taking into account both real human motion capture data as well as accurate geometric information of the dynamic 3D scene. With different lighting and environment conditions, we have generated various shaking videos that match the different motion-types such as jogging, walking, running, etc. As a result, we believe that our shaking video synthesis method opens the door to better objective evaluation of video stabilization algorithms. The proposed framework's flexibility and accuracy will benefit and ease future research on 3D-aware video stabilization.

2. MOTION CAPTURE DATA BASED CAMERA TRAJECTORY SYNTHESIS

We aim to synthesize a camera motion trajectory when using real-world motion measurements. That is, in order to simulate shaking video captured by a moving user in the 3D scene, it is preferable to use real human motion-data rather than a motion model. While it is relatively easy to obtain recorded 6 DOF motion data, there are, however, several other issues that still need to be addressed. First of all, the motion trajectories in the motion capture dataset are relatively short in time. Moreover, it is highly unlikely to find an existing path in the dataset that directly matches the desired camera path specified by the user. Therefore, we need to process the desired path, adding fine details of matching motion capture data in an external dataset. Now we describe how we synthesize the details of the motion data and integrate it into the desired camera trajectory.

2.1. Motion Capture Data and Its Extension

The original shaking motion-data we adopted in our work is from the human motion activity database [12], named CMU-MMAC, which is collected by Carnegie Mellon University's Motion Capture Lab. The dataset was captured by a motion capture system with multi-modal sensors including twelve infrared cameras, five 3-axis accelerometers and gyroscopes. This dataset contains measurements for motion-types such as walking, running, dancing, climbing and various other human behaviors. In our system, the 3D translation and 3D rotation information provided in this dataset are of critical importance for our camera motion simulation. Fig. 2(a) and 2(b) show such 6 DOF motion information. Notice that the translation on the z-axis represents forward/backward motion.

Because most of the motion segments in the CMU-MMAC dataset contain around 400 samples at 120Hz, all segments are rather short (many of them are only $3s$). Therefore, we need to temporally-extend the data to match the relatively long-term camera motion in our system. As shown in Fig. 2, the CMU-MMAC motion data is characterized as periodic. This is easy to understand. For instance, the human walk activity itself can be regarded as periodic behavior. In this context, we extract all the motion periods of the motion data, and for each motion category we construct thousands of samples by randomly combining those segments having different periods. It is important to notice that such data combination could introduce gaps or errors between segments. Among them, the most significant is the drift error, which also widely occurred

Fig. 3. We rendered various dynamic 3D scenes for shaking video synthesis. The system allows the user to flexibly configure the scene with day/night lighting, simple/complex textures, and static/moving foreground objects according to the user's preference.

in other 3D-oriented trajectory reconstruction problems [13]. In our system, we apply a high-pass filter, i.e. the Butterworth Filter [14], to eliminate the drift error.

Next, we synthesize the motion data using frequency-based techniques, and synthesize the desired camera trajectory.

2.2. Camera Trajectory Synthesis

Our purpose is to apply external motion capture data to the camera motion in our 3D scene modeling. However, those paths in the dataset are diverse and they may be not matched to the specified camera trajectory. Fortunately, the motion capture data shows strong periodic behavior. Therefore, it is intuitive to construct Fourier-based approaches to efficiently reuse existing signals. A similar idea using frequency-based analysis has been applied in [2], where the authors used it to reconstruct the parameters of a homography matrix for warping a 2D video. In our solution we employ the Inverse Discrete Fourier Transform (IDFT), as it is still suitable even when the input motion data is aperiodic. Formally, the analysis function in our solution is:

$$p[t] = \frac{s}{\sqrt{n}} \sum_{i=1}^{n} f[i] e^{j2\pi i} + n[t], \tag{1}$$

where $p[t]$ is the desired shaking path at a certain time instance t, s is a scaling factor which will compensate the lost energy, $f[i]$ is the frequency component of sample i, $n[t]$ is an additive noise component at instance t, and j denotes the imaginary unit, $j^2 = -1$.

When observing the spectrum of motion data (see one example in Fig. 2(c)), it is easy to notice that most of the spectrum energy only contains few dominant frequency components that are localized between 0Hz to 20Hz. This allows us to extract the probability distribution of the spectrum by focusing on these dominant components. Thus, we obtain the first 10 dominant frequency components by searching the local maxima of the spectrum, and use a Gaussian distribution function to fit the frequency, phase and intensity of every local maximum component. Once we have estimated the mean and variance of each Gaussian, we can reconstruct a spectrum $f[i]$ that contains the dominant frequency components using the following equation:

$$f[i] = re^{j\theta}, \tag{2}$$

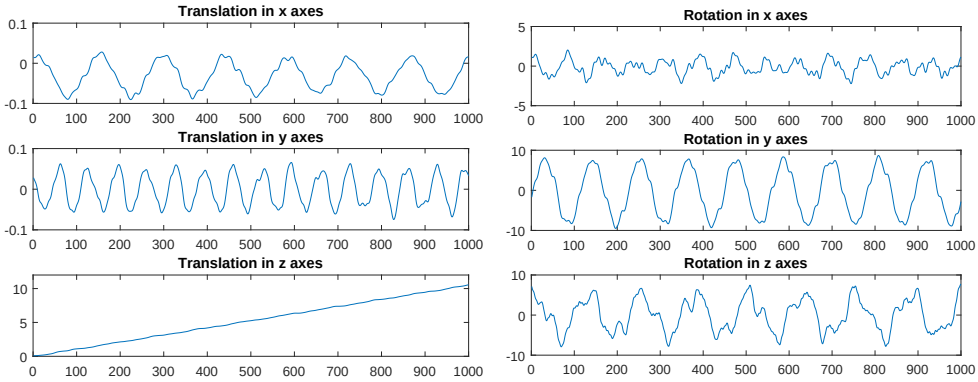
where $i = \frac{nF}{f_s}$. The parameters r, θ and F are the intensity, phase and frequency factors, respectively. .Note that they are generated by a Gaussian distribution generator based on the real motion data. f_s is the sampling frequency, and n is the number of samples of the motion capture data.

By substituting $f[i]$ into Eq.(1), we can compute the shaking video trajectory in the time domain. This shaking data can then be mapped into any arbitrary motion path. Therefore, we take such shaking data as high-frequency information and integrate it into the expected motion path. The generated camera trajectory data is finally imported into the dynamic 3D scene (see examples in Fig. 5). The noise component $n[t]$, modeled with different Signal-to-Noise Ratios (SNRs), could be easily introduced into the synthesized trajectory data.

3. DYNAMIC 3D SCENES

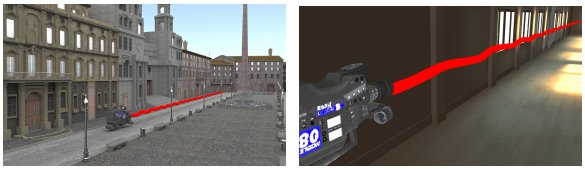
The proposed shaking video synthesis framework provides a system to model dynamic 3D scenes. The 3D render system allows the user to flexibly configure the scene with day/night lighting, simple/complex textures, and static/moving foreground objects according to the user's preference. Moreover, the camera motion trajectory generated by our approach is automatically imported into the 3D rendering engine.

The proposed shaking video rendering system also has an interface for the user to assign both the degree-of-shaking and a coarse motion path of the camera in the 3D scene. The user can further define what type of shaking motion is required in the output video by choosing different kinds of motion categories (walking, jogging, running, etc.) and shaking noise distributions.



(a) Translation components. (b) Rotation components.

Fig. 4. Synthesized camera trajectory using the proposed frequency-based extension. The generated camera motion trajectory shares explicit periodic characteristic with the original motion capture data, while effectively avoiding drifting errors or gaps.



(a) (b)

Fig. 5. Our synthesized camera trajectories in 3D scenes.

4. EXPERIMENTS AND DISCUSSIONS

The proposed system is implemented on a Windows10 desktop with Intel Core i7-4790K CPU, 32GB memory and 2 Titan X graphics cards. The camera motion trajectory simulation is implemented in Matlab 2017, the dynamic 3D scene and corresponding videos are rendered by Autodesk Maya. The system control framework as well as the communication functions are developed in Python. Our system synthesizes the camera trajectory with real-time interactive feedback, and the run-time mainly depends on rendering video frames. In general, for full HD resolution video, it takes about 1.5 seconds to render each frame.

We generated various shaking and stable videos in different 3D scenes (see the supplemental demo). As shown in Fig. 3, the proposed system provides complex lighting environments with sunlight, reflections of the background, or moonlight at night. The user can define a simple or complex background for the target 3D scene. Besides, motion objects, such as a dancing human or a moving car/bus, could be easily integrated into the synthesized video. One of the synthesized camera trajectories is shown in Fig. 4. The generated trajectory shares explicit periodic characteristic with the original motion capture data, and it follows the motion rhythm as we expected. Moreover, the result effectively avoids drift errors or gaps between different periods.

When evaluating video stabilization, it is unsuitable to run some well-studied 2D-to-3D matching (e.g. [15, 16]) or feature-based indexing approaches (e.g. [17]). As every captured video frame is different in the dynamic 3D scene, it is difficult to organize and retrieve the data for all point clouds of the dynamic 3D scene. It has been pointed out that the stabilized motion trajectory should be as close as possible to the original camera trajectory [18]. Furthermore, when performing stereo video stabilization [19], the trajectory should be further considered. We will thus investigate the trajectory produced by the stabilization method, and check how far is the stabilized trajectory from the ground-truth. Our system is also flexible to import other motion capture datasets and 3D scenes. Moreover, the proposed framework can render videos with some special effects. For instance, when the the point spread function [20, 21], exposures and the noise are defined, blurred videos could be synthesized; this allows for further comparative evaluation of view synthesis [22, 23, 24, 25] or deblurring algorithms [26] in 3D environments.

5. CONCLUSIONS AND FUTURE WORK

This paper presents a shaking video synthesis approach by combining human motion capture data with man-made dynamic 3D scenes. The proposed framework is capable to provide both the accurate 3D geometric information and the shaken motion information in detail, allowing to obtain the ground-truth motion and video pairs and enabling the comparative assessment of video stabilization algorithms. In the future, we will investigate the performance of existing video stabilization algorithms by evaluating not only the stabilized camera trajectory but also the corresponding geometrical distortion of the 3D scene.

6. REFERENCES

[1] Feng Liu, Michael Gleicher, Hailin Jin, and Aseem Agarwala, "Content-preserving warps for 3d video stabilization," *ACM Trans. Graphics*, vol. 28, pp. 44, 2009.

[2] Hui Qu, Li Song, and Gengjian Xue, "Shaking video synthesis for video stabilization performance assessment," in *Proc. VCIP*, 2013, pp. 1–6.

[3] Heng Guo, Shuaicheng Liu, Tong He, Shuyuan Zhu, Bing Zeng, and Moncef Gabbouj, "Joint video stitching and stabilization from moving cameras," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5491–5503, 2016.

[4] Miao Wang, Guo-Ye Yang, Jin-Kun Lin, Ariel Shamir, Song-Hai Zhang, Shao-Ping Lu, and Shi-Min Hu, "Deep online video stabilization," *arXiv preprint arXiv:1802.08091*, 2018.

[5] Lei Zhang, Qing-Zhuo Zheng, and Hua Huang, "Intrinsic motion stability assessment for video stabilization," *Preprint in IEEE Trans. Vis. Comput. Graphics*, 2018.

[6] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graphics*, vol. 36, no. 4, 2017.

[7] "Motion analysis corporation," `https://www.motionanalysis.com/`, Accessed: 2017-06-20.

[8] "Vicon motion systems," `https://www.vicon.com/`, Accessed: 2017-10-08.

[9] "Polhemus, inc.," `http://polhemus.com/`, Accessed: 2017-12-30.

[10] Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović, "Practical motion capture in everyday surroundings," *ACM Trans. Graphics*, vol. 26, no. 3, pp. 35–1, 2007.

[11] Marek Vondrak, Leonid Sigal, Jessica Hodgins, and Odest Jenkins, "Video-based 3d motion capture through biped control," *ACM Trans. Graphics*, vol. 31, no. 4, pp. 27, 2012.

[12] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran, "Guide to the carnegie mellon university multimodal activity (CMU-MMAC) database," *Robotics Institute*, p. 135, 2008.

[13] Parvaneh Saeedi, Peter D Lawrence, and David G Lowe, "Vision-based 3-d trajectory tracking for unknown environments," *IEEE Trans. Robotics*, vol. 22, no. 1, pp. 119–136, 2006.

[14] Ivan W Selesnick and C Sidney Burrus, "Generalized digital butterworth filter design," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1688–1694, 1998.

[15] Torsten Sattler, Bastian Leibe, and Leif Kobbelt, "Fast image-based localization using direct 2D-to-3D matching," in *Proc. ICCV*, 2011, pp. 667–674.

[16] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt, "Scalable 6-DOF localization on mobile devices," in *Proc. ECCV*, 2014, pp. 268–283.

[17] Youji Feng, Lixin Fan, and Yihong Wu, "Fast localization in large-scale environments using supervised indexing of binary features," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 343–358, 2016.

[18] Hui Qu and Li Song, "Video stabilization with L1-L2 optimization," in *Proc. ICIP*, 2013, pp. 29–33.

[19] Heng Guo, Shuaicheng Liu, Shuyuan Zhu, and Bing Zeng, "Joint bundled camera paths for stereoscopic video stabilization," in *Proc. ICIP*, 2016, pp. 1071–1075.

[20] Curtis R Vogel and Mary E Oman, "Fast, robust total variation-based reconstruction of noisy, blurred images," *IEEE Trans. Image Process.*, vol. 7, no. 6, pp. 813–824, 1998.

[21] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, 2011.

[22] Shaoping Lu, Jan Hanca, Adrian Munteanu, and Peter Schelkens, "Depth-based view synthesis using pixel-level image inpainting," in *Proc. DSP*. IEEE, 2013, pp. 1–6.

[23] Shao-Ping Lu, Beerend Ceulemans, Adrian Munteanu, and Peter Schelkens, "Spatio-temporally consistent color and structure optimization for multiview video color correction," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 577–590, 2015.

[24] Beerend Ceulemans, Shao-Ping Lu, Gauthier Lafruit, Peter Schelkens, and Adrian Munteanu, "Efficient mrf-based disocclusion inpainting in multiview video," in *Proc. ICME*. IEEE, 2016, pp. 1–6.

[25] Beerend Ceulemans, Shao-Ping Lu, Gauthier Lafruit, and Adrian Munteanu, "Robust multiview synthesis for wide-baseline camera arrays," *Preprint in IEEE Trans. Multimedia*, 2018.

[26] Michal Šorel and Filip Šroubek, "Space-variant deblurring using one blurred and one underexposed image," in *Proc. ICIP*. IEEE, 2009, pp. 157–160.